Personal Semantics

Gregory Grefenstette

Personal Semantics

Gregory Grefenstette

Abstract Quantified self, life logging, digital eyeglasses, technology is advancing rapidly to a point where people can gather masses of data about their own persons and their own life. Large-scale models of what people are doing are being built by credit companies, advertising agencies, and national security agencies, using digital traces that people leave behind them. How can individuals exploit their own data for their own benefit? With this mass of personal data, we will need to induce personal semantic dimensions to sift data and find what is meaningful to each individual. In this chapter, we present semantic dimensions, made by experts, and by crowds. We show the type of information that individuals will have access to once lifelogging becomes common, and we will sketch what personal semantic dimensions might look like.

1 Introduction

Extracting and finding information in large quantities of unstructured data requires assigning data into semantic classes, so that information can be filtered merged, and labelled. In applied Natural Language Processing, experts produced validated taxonomies and ontologies, such as the Medical Subject Heading (MeSH) and the NASA thesaurus for classifying text. With the expansion of the internet and Web 2.0, new, crowdsourced knowledge structures began to appear, for example, the DMOZ hierarchy of the Open Directory Project, and the category hierarchy of Wikipedia. Both formal and crowdsourced taxonomies allow semantic annotation of information, and are used to accelerate search, allowing the user to choose categories and other metadata before examining results. In a sense, the categories used are an agreed-upon

Gregory Grefenstette
INRIA, Saclay France, e-mail: gregory.grefenstette@inria.fr

(either by experts or the crowd) way of looking at the world, and classifying things in it. I believe we will soon need a new third type of semantics, a personal semantics that can be automatically generated but limited to one person's view of the world. This new type of semantics will be needed to organize the digital traces that individuals create. In the near future, due to advances in video and audio processing, in GPS tracking and in memory storage, people will be able to record their lives unobtrusively in video, audio, and position. Buttonhole audio cameras mentioned, devices such as Google Glass, will soon be cheaply available to the general public. Each person will be generating personal multimodal Big Data about their lives. In order to process (index and retrieve) this multimodal data, it will have to be semantically annotated in an automatic fashion, since people will not have the time to manually review their own captured date. Contrary to the first two types of shared semantics (expert, crowdsourced), each person will have personal semantic categories, places, people, events and other categories that meaningful to them alone. This chapter will examine how personal semantics can be generated from the personal data gathered from portable lifelogging devices, mixed with digital traces, and open linked data.

2 Semantic Dimensions

Semantic dimensions help people find things faster. In modern search engines, semantic dimensions are materialised as *facets*. Facets are usually presented as a list of expandable categories. For example, if you type "wool" on a popular shopping website, you see the following "Departments" on the left side of the screen: *Arts & Crafts, Clothing & Accessories, Books, Home & Kitchen, ...* . Each of these "Departments" is a facet, or semantic dimension, that allows you to divide the universe into two parts: things that belong to that dimension and things that do not belong. Expanding one of the department facets opens up further subdivisions. For example, click on *Clothing & Accessories* and the shopping site displays the narrower facets: *Women, Men, Accessories, Novelty & Special Use, Boys, Girls, Baby, Luggage, Handbags*. When we shop, these subdivisions seem natural, or at least immediately comprehensible, even though, of course, the semantic categories *Boys*, *Girls* and *Baby* are not mutually exclusive[1]. The shopping user is not bothered by the lack of formality, and naturally assumes that the dimension *Boys* covers clothings designed for male children somewhere between roughly the ages of 2 and 17, with *Baby* being the dimension of younger humans and *Men* for older male humans.

[1] When you click on *Baby* on this site, you find the following three dimensions: *Baby Boys*, *Baby Girls* and *Unisex*.

Other semantic dimensions that appear under *Clothing & Accessories > Baby* on this same web site are *Price* (with intervals such as *$25 to $50*), *Brand*, *Size*, *Average Customer Review*.

Common users who shop online are now used to using facets to focus in on the items they are searching for. Naturally and intuitively, they have learned over the past decade to combine the query box method of search as they use on Google, with the use of facets to restrict or refine their search while they shop.

Business users, using modern enterprise search systems also use facets to speed retrieval. An enterprise search will index the documents within an entreprise using keyword indexing, but also using the metadata associated with business documents to create facets. In this case, common facets are type of document (Word, PDF, Powerpoint, email, posting), the date of production of the document, the business department that the document is attached to (Markering, Communication, Customer Relations, Management, etc.), the creator of the document, the sender and receivers, the language, and the product and the people named in the document. Most of this information is not considered semantic information by the natural language processing community, but users use these facets in exactly the same way as in the case of shopping facets, to reduce the space of search. These facets can be considered semantic dimensions since they associated typed information with the document, information that is not necessarily found in the keyword-indexed text of the document.

In general, where do semantic dimensions come from? In the rest of this chapter, we will examine three different ways of creating semantic dimensions: via experts, via crowdsourcing, and via induction from data.

3 Expert Semantic Resources

Every since collections of writings have existed, there has been a need of knowing how to order the collection. In libraries of printed books, this physical need to be in one place and desire to group books about the same subject together gave rise to library classification schemes, such as the Library of Congress Classification (1897), the Dewey Decimal system (1876), etc.

Before computers, search was performed using printed catalogs. There was a real cost, associated with the paper it was written on, of including a piece of information in the catalog. The constraints of space and cost naturally led to controlling the indexing language, which led to "authority lists" of the categories and subcategories which could be associated with a piece of information.

WordNet is another example of organizing concepts (at least single word concepts) in a sematnic hierarchy. To provide a resource for analyzing psychological text, Miller and his team collected definitions from a number of

Class 000	Computer science, information & general works
Class 100	Philosophy and psychology
Class 200	Religion
Class 300	Social sciences
Class 400	Language
Class 500	Science
Class 600	Technology
Class 700	Arts & recreation
Class 800	Literature
Class 900	History & geography

Fig. 1 The Dewey Decimal System is an long-used classification system for libraries. It divides subjets into a hierarchy, with the uppermost classes shown here. Still in use, the lower nodes of the hierarchy are modified once a month (http://oclc.org/dewey/updates.en.html)

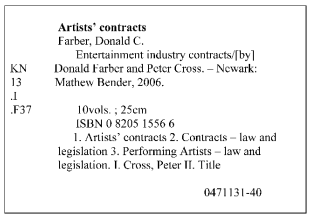


Fig. 2 Supplementing the one-book, one-place paradigm, printed card catalogs allowed a book to be indexed under different dimensions (author, title, subjects).

dictionaries, and arranged words in a hierarchy of synsets (a synset is a set of synonyms). Dictionary definitions are often of the the structure: A is a type of B in which C, where A is the head word, B is a more general class and C are the differentiators that distinguish C from other elements of the class B. B is called the hypernym of A and A is called a hyponym of B. WordNet is a hierarchy of hypernyms and hyponyms over words (including a few proper nouns) of English. An ambiguous word can be found in many synsets, but each individual meaning is found in only one place in the hierarchy.

Shiyali Ramamrita Ranganathan, developed the Colon Theory of Classification in the early 1930's. It was widely adopted by libraries afterwords. This colon based notation assigned different semantic classes to each work, separated by colons, whence the name. Each hierarchical class corresponds to facet in modern information retrieval.

Beyond general classification of human knowledge, domain-specific classifications also began to appear in the 1950s. The National Library of Medicine in the US, first developed a Subject Heading Authority List in 1954, that evolved over time into the Medical Subject Headings (MeSH). MeSH headings are used to index the more than 20 million medical articles appearing in the bibilographic database MedLine (https://www.ncbi.nlm.nih.gov/pubmed). It

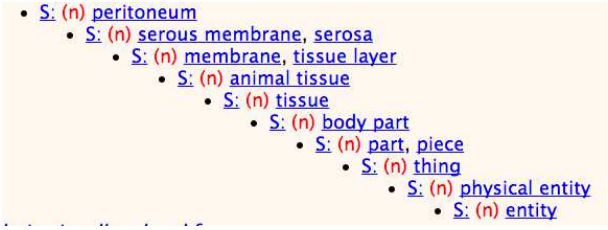


Fig. 3 A slice of the WordNet hierachy, showing the hypernyms (more general classes) for the word *peritoneum* which is a *serous membrane, or serosa* which is a *membrane, or tissue layer* which is an *animal tissue* which is a *tissue body part* which is a *part or piece* which is a *thing* which is a *physical entity* which is a *entity*. These are all nouns (indicated by the *n*). The *S* stands for synset. Wordnet also contains verbs, adjectives and adverbs.

is updated regularly by a permanent staff of ten doctors, veterinarians, and PhDs at the National Library of Medicine in the US.

Fig. 4 A sample of the MeSH semantic hierarchy. The Abdomen (A01.923.047) is part of the Torso (A01.923) which is a Body Region (A01). The category A concerns Anatomy (of humans and animals). Other categories are B for Organisms, C for Diseases, D for Chemicals and Drugs, etc. (See http://www.nlm.nih.gov/mesh)

In the 1960s, the National Aeronautics and Space Administration (NASA) produced its first thesaurus. It is updated monthly by engineers and lexicographers in the NASA Scientific and Technical Information program http://www.sti.nasa.gov/about-us/. The thesaurus contains over 18,000 terms in the fields of aeronautics and engineering. This semantic resource has been used for automatically annotating new documents since at least 1994 (6).

The MeSH and NASA thesaurus are examples of expert-directed semantic structuring of a domain. They are expensive to maintain, updated monthly by commitee decision, and directed to an audience of specialists.

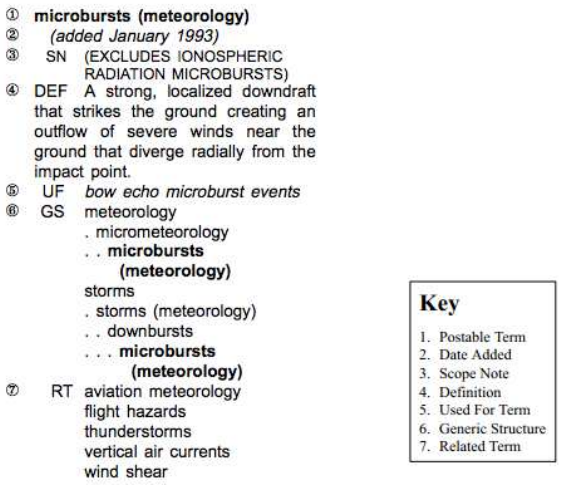


Fig. 5 A typical hierarchical entry in the Nasa thesaurus. *Microbursts* appear under two facets: Meteorology, and Storms. (See http://www.sti.nasa.gov/thesvol1.pdf)

4 Crowd-sourced Semantic Hierarchies

In contrast to expert design and maintained semantic structures, we have seen crowd-sourced semantic hierarchies developed over the past twenty years. Crowd-sourcing here means that a large number of "ordinary"' people, for example, web users, can contribute and alter entries in the semantic hierarchy.

One of the first crowd-sourced efforts to structure the information on the Web was the Open Directory Project (ODP , at dmoz.org), begun in 1998 by two SUN engineers, Rich Skrenta and Bob Truel, beginning with a hand-built hierarchy dervied from USENET news groups. The orignal idea here was to allow anyone on the internet to become a directory editor, after proving their ability to correctly edit a small portion of Web pages. This open community-created semantic hierarchy is used by a number of other search engines: Netscape Search, AOL Search, Alexa, and Google (until 2011).

Inspired in part by ODP, Wikipedia created an open source encyclopedia, allowing anyone to create and edit pages, depending on the crowd to police edits, removing errors and spam (certain offending IP addresses can be banned). Wikipedia pages can be categorized. Categories are also crowd-sourced. For example, in the German version of Wikipedia, the article for *Fersental* (in English, the Mocheni Valley) was categorized as in the following categories: *Sprachinsel* (Isolated langauges), *Tal im Trentino* (Valleys in Trento), *Deutscher Dialekt* (German dialects). Some categories are listed in Wikipedia as subcategories of other categories. for example, *Valleys in Trento* is a subcategory of *Valleys of Italy*. Gerard de Melo and Gerhard Weikum

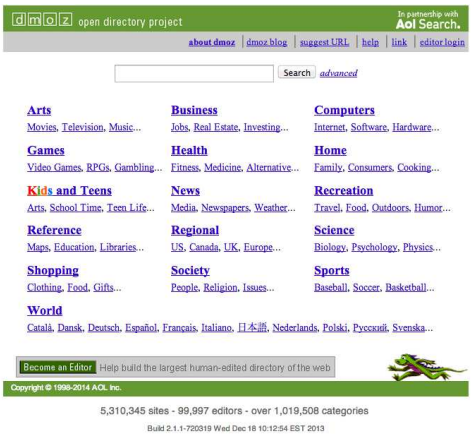


Fig. 6 The front page of dmoz.org. There are over 5 million web pages hierachically indexed in over 1 milllion categories by almost 100,000 editors.

described how this graph, which extends over language versions of Wikipedia can be structured into a semantic hierarchy (4).

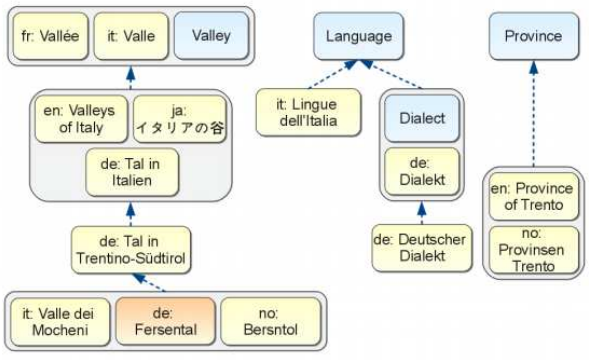


Fig. 7 A multilingual semantic hierarchy induced from the Wikipedia category graph, by Gerard de Melo and Gerhard Weikum into METANET (4)

These two semantic hierarchies are controlled by the crowd, which collectively decides what should appear and what shoud not, the consent of the hierachies resulting from consensus over shared community viewpoints.

5 Personal Hierarchies

The semantic hierarchies presented categorize public information into classes that anyone from the user community can understand. Public information, public semantics.

Computer technology in wearable and portable computing has reached a point where individuals are able to passively gather large quantities of information about themselves, stored in digital format. As this information grows in size and storage costs continue to drop, it is probable that people will collect their own life logs, with the data that the user generates and interacts with. In this mass of data, individuals will have the same need for classifying information, grouping information into classes, so that search and retrieval can be efficiently performed. The difference with systems developed for public data, is that the semantic classes used need only make sense to the individual. This information is his or her private information, and they may order it in any way they like, without having to explain.

Certainly some dimensions may be comprehensible to others, but this will not be a necessity.

Above all, it will be necessary to automatically create many of these semantic dimensions, and to automatically annotate the data that the user generates. It is enough to live one life, without having spend another life annotating and classifying it.

5.1 Personal Data Sources

Here we will look at some of the data sources that people generate or will soon generate in their daily lives.

5.1.1 Text and Browsing Behavior

People who use computers and communication devices generate a lot of text: emails, text messages, posting in social networks, chats, local computer files. They also attract a lot of information to them: email received, messages posted by others on their personal walls (such as on Facebook or Google+), content of web pages that they browse.

Currently much of this information is exploited by third parties (advertisers, national security agencies) because this text reveals some of the personality of the user, what interests them, what they might want to do. For example, Google and other companies offer free e-mail services to user. In exchange, Google will analyze the contents of your email, in order to "sell" space on your screen to advertisers in function of that email content. Similarly, many web pages or web servers introduce invisible 1x1 pixel images

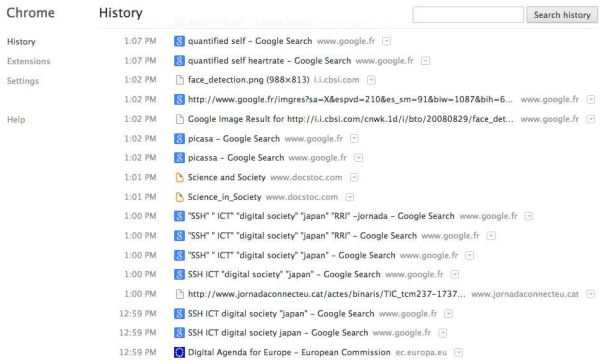


Fig. 8 Search engines, such as Google, keep a history of the web sites you visit. They keep an index of the titles of the pages viewed, along with a timestamp. This information can model user interests.

into web pages served to users. When the web page is displayed, a request for the image is sent to the web site that owns the 1x1 pixel image, along with the URL of the requesting page and your IP address. When this image is owned by an advertising agent, they learn from analyzing the content of the web page (from its URL address) and from your IP address (associated with your computer), what subjects you are looking at, and from this, these agents can create a model of your interests that can be used to serve ads. This is why, after looking for a vacation hotel on a travel website, you can see ads appearing for other hotels in that locality on completely different websites. Advertisers do a semantic analysis of your behaviour to serve targeted ads. Browser add-ons such as Ghostery can block this traffic.

This information can be used for your own benefit, too, if you capture it and analyze it. In a business setting, browsing behavior of employees can be used to identify expertise within an organization. In US Patent 6,446,035, a system is described that stores and analyzes web pages browsed by workers, browsing in work mode, in an organization. The textual content of the page is reduced to normalized noun phrases, and categorized, and stored along with the bowser's identity. When someone else is searching for someone knowledgable about some area within their organisation, they can search this stored information for people who have also browsed that topic.

Currently, people do not use their own browsing behavior to create any stored version of their own interests, relying on actively set bookmarks, or simple search histories containing only the title of the pages visited, to keep a record on things that they have browsed. Searching though email, chats and posting is also limited to simple string search.

There have been many research attempts to automatically classify emails, given an existing set of directories, or into two classes of spam or not-spam (8), or without classifying into directory using topic-detection techniques (2).

5.1.2 Wearable Computing and Quantified Self

In addition to explicitly written text, soon people will be generating a lot of data from devices that they carry with them. The most ubiqitous example of such wearable computing is a person's cell phone, which interacts with communication towers to pinpoint the user's location, in order to receive and send calls. This information is exploited by security agencies for tracking "persons of interest", and can be exploited by a user via a number of tracking apps that they can install on their cell phone. For example, My Running Pal will track a bicycle route, or a run, that can then be sent to another user or social website.

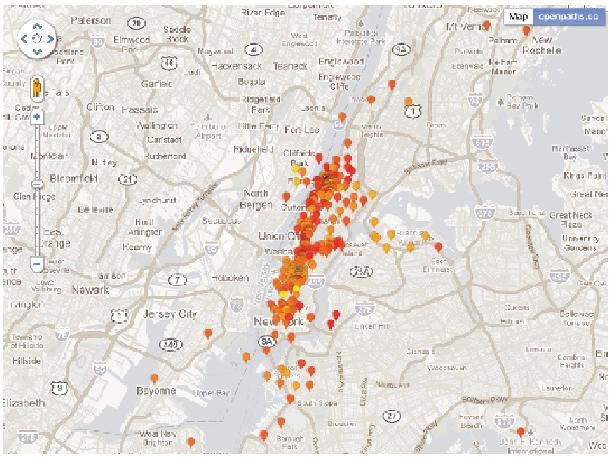


Fig. 9 Passive location tracking applications, such as OpenPath, can capture GPS coordinates from a cell phone, and display personal position during a given time period. Repeated paths and locations can be used to identify common routes taken by the wearer. Information along these routes could be gathered from open business directories to help the user conveniently find items on a shopping list, for example.

OpenPaths is an application that keeps track of your GPS coordinates, sends them to a central repository where the owner can access them, or allow an approved third-party to download an anonymized version for scientific study.

An application of such tracking for a user's own benefit could be the following scenario. From the timestamped GPS information in OpenPaths, it is possible to detect common routes taken by the wearer. These routes can be crossed with information about stores along these routes, using open source maps such as OpenStreetMap. This crossed information would reveal to a user where they could buy an item that they need, by indicating stores that might carry it along their daily path.

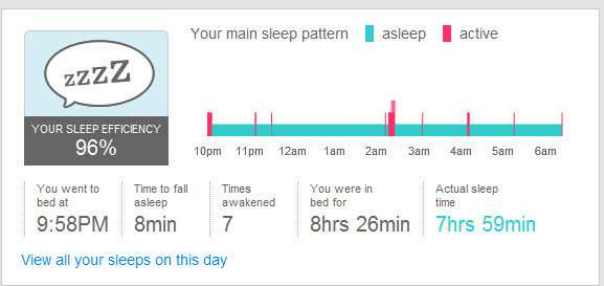


Fig. 10 One output of the wearable quantified self bracelet, FitBit, is an image of how the wearer slept, measuring periods of calm and activity during the night.

In addition to tracking GPS information, wearable computing can capture a range of other data: a person's heart rate, the number of steps taken in a certain period, temperature, blood pressure, andother vital signs. A movement called Quantified Self, of people who are tracking this data about themselves, for their own use or to share, has led to great number of commercial products: FitBit, FuelBand, Jawbone Up, Basis Band, Cardiio, Azumio, Beyobe, ...

These time-stamped personal data can be mapped onto emotional and physical states: calm, happy, agitated, active, sleeping, ill, ..., that are specific to a given person, and which that person might use to find and retrieve certain events in their logged lives(7).

5.1.3 Digital Eye Glass

In 2012, Google announced Google Glass, a prototype of a wearable video and sound capture device, set to be commercialized in 2014. By 2013, a great number of competitors have appeared (Telepathy One, Sony Smart Glass, Microsoft Augmented Reality, Vusix, ReconJet, MetaSpace Glasses, Oakley, GlassUp, Oculon Electronics, OptiVent, Epiphany Eyewear, castAR, 13th Lab,). Mann has been wearing and developing digital eye glasses since the 1980s. Though currently constrained by battery life, soon these glasses will allow their user to record their entire days, in video, image and sound.

This data can be converted into symbolic, storable data through video and image processing, and through speech to text analysis. An example of personal semantics applied to image processing is the family-and-friend recognition that was made available in Picassa in the early 2010s. In this photo processing system, all your locally stored photos were analyzed to identify faces (essentially ovals with two eyes) and these faces were clustered. Picassa would then present you with an interface in which you could associate a name with tight, precise clusters, remove faces if need be. With this cleaned and labeled information, Picassa would create a model of each named person in

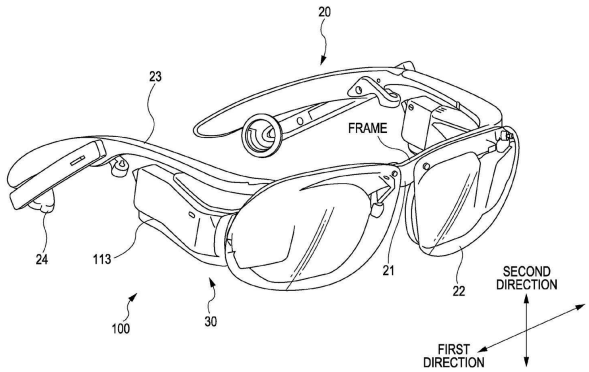


Fig. 11 Google Glass has many competitors appearing. Here is an image from a Sony patent application for a similar wearable digital eyeglass. Sony has also filed a patent, US 20130311132, for a smart-wig that embeds sensors inside a false hairpiece.

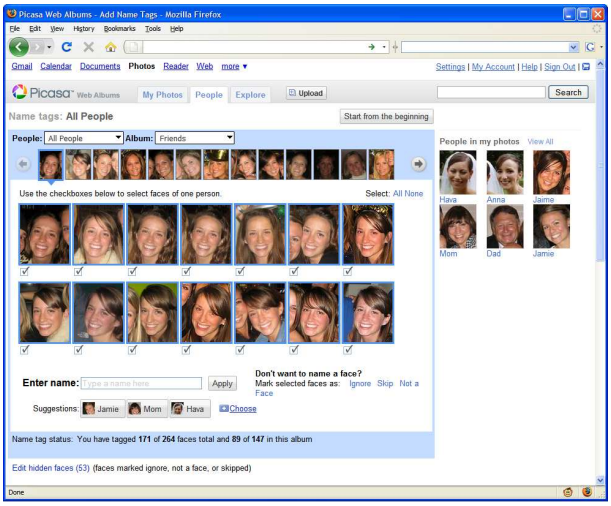


Fig. 12 Picasa, now part of Google Images, introduced a level of personal semantics by creating personal models of faces found in one user's photographs. These faces were clustered and the cluster was presented to the user for editing and for labeling. Newly added labeled images were used to reinforce the model of the labeled face.

your local photos, that would be used to identify less evident faces. Facebook has also adopted a similar software.

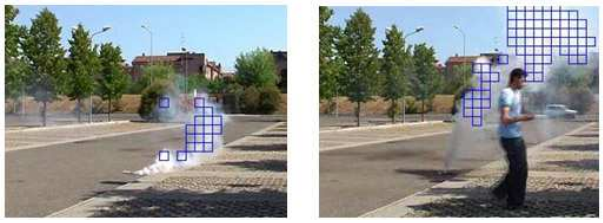


Fig. 13 Video processing can currently recognize a number of events in video: cars and people moving, fire, crowds, etc. Here in a system produced for the TrecVid by the Informatics and Telematics Institute, Centre for Research and Technology Hellas (CERTH-ITI), *smoke* is automatically detected, even when someone with a white shirt passes in front of it. (1)

Video processing is currently limited to a few large classes: detecting crowds, movement of people, smoke/fire, indoor, outdoor, etc. But as the number of classes grow these classes could also be used to annotate personal data. Identifying persons through face identification is well advanced.

Speech-to-text processing is getting better all the time (3). Since 2012, Apple has included the speech recogntion application SIRI in its iPhone offer. Spoken questions are transmitted to a distant server for recognition, but the recognition is biased towards items found in the user's local context: contact names, physical location, time of day.

Fig. 14 Siri is currently the most popular speech-to-text application. As this technology continues to improve, it will be possible to have wearable computing passively capture speech throughout the user's day.

5.2 Sketch for Personal Semantic Dimensions

It is easy to project that, in the near future, people will have access to the following information passively collected about their own lives:

- their GPS position at any given moment
- all the things that are around those GPS points
- their vital signs at any given moment
- all the emails they have written
- all the webpages they have browsed
- all the chats, text messages, phone calls they participated in
- all the mails and messages received
- all the things they have seen or read
- all the words that they have said or heard
- all the noises they have hear

As in a cluttered house, it will be difficult to find things in this data without some type of organisation. To retrieve some piece of information, we can imagine that 'the mental dictionary is a huge semantic network composed of words (nodes) and associations (links)" that can lead us to the information we want (9). Some of these links will be socially shared and we can assume that the associations between items are those found in socially shared views of the world, such as are found in the semantic resources prepared by experts (MeSH, WordNet), or those created by crowdsourcing (DMOZ, Wikipedia categories). But other associations and links will depend on the person's own mental lexicon, on what places and objects mean to the person, on who the person knows and cares about, and why.

From psychology research on personal semantics, we see that people structure their memories in terms of autobiographical facts (facts about their own lives), episodic memories (repeated or unique events), general knowledge, and autobiographically significant events linking events to general events in the world (5). In addition to this structuring, people can have procedural knowledge, lexical knowledge, and certain brain injuries can affect one of these memories structures and not the others, so one can remember how to use a machine but not remember any of the words for the individual parts of the machine. One can remember personal events and lose track of general knowledge, or the contrary.

To structure the passively collected information, we will need to apply expert semantic hierarchies, crowd-sourced hierarchies and hierarchies induced by techniques, yet to be determined, to the user's personal passively collected data. Adding annotations from these hierarchies will provide the association links into one's personal data.

A rough example of these hierarchies might be the following. Suppose that one searches in one's personal archives for "wool", this query might procude the following search facets (in addition to presenting snippets and thumbnails of top matches):

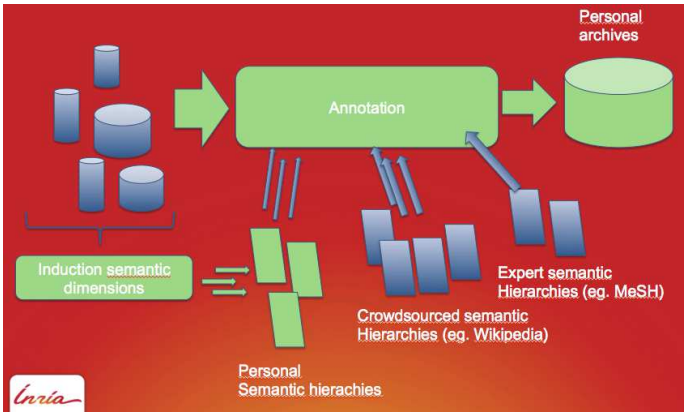
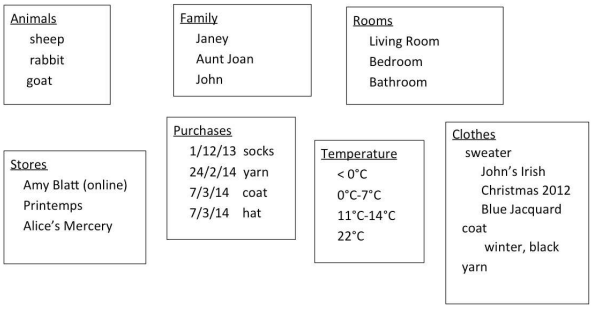


Fig. 15 From many different personal information sources shown on the left of this diagram, we will have to use image, speech, video and text processing to extract personal semantic dimensions. These dimensions as well as crowdsourced semantic hierarchies, and expert defined hierarchies can then be used to annotate a user's personal data.



Clicking on a facet here would select items in the personal archive that are annotated with the part of the mutliple hierarchies. These items could be images, videos, events, places on a map, records of purchases, documents, or any of the other types of data captured in personal archive.

6 Conclusion

Technology has reached a point where it will soon be feasible to capture, store and process great portions of people's lives, for good or for bad. One may always chose not to record their life, as one may choose not to use a cell phone, or credit cards today. I believe that the advantages of having traces of our lives will outweigh the drawbacks. I am not sure what exactly the personal semantic dimensions will be, will they resemble each other from

person to person, or be completely incomprehensible to another person? I believe we will soon see, because in the mass of information that we can collect, only categorization will allow rapid search and retrieval. And these categories must make sense to the individual.

References

[1] Avgerinakis, K., Briassouli, A., Kompatsiaris, I.: Activity detection and recognition of daily living events. In: Proceedings of the 1st ACM international workshop on Multimedia indexing and information retrieval for healthcare, pp. 3–10. ACM (2013)

[2] Cselle, G., Albrecht, K., Wattenhofer, R.: Buzztrack: topic detection and tracking in email. In: Proceedings of the 12th international conference on Intelligent user interfaces, pp. 190–197. ACM (2007)

[3] Lamel, L., Courcinous, S., Despres, J., Gauvain, J.L., Josse, Y., Kilgour, K., Kraft, F., Le, V.B., Ney, H., Nußbaum-Thom, M., et al.: Speech recognition for machine translation in quaero. In: Proceedings of the International Workshop on Spoken Language Translation (IWSLT), San Francisco, CA (2011)

[4] de Melo, G., Weikum, G.: Menta: inducing multilingual taxonomies from wikipedia. In: Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 1099–1108. ACM (2010)

[5] Renoult, L., Davidson, P.S., Palombo, D.J., Moscovitch, M., Levine, B.: Personal semantics: at the crossroads of semantic and episodic memory. Trends in cognitive sciences **16**(11), 550–558 (2012)

[6] Silvester, J.P., Genuardi, M.T., Klingbiel, P.H.: Machine-aided indexing at nasa. Information Processing & Management **30**(5), 631–645 (1994)

[7] Swan, M.: Next-generation personal genomic studies: extending social intelligence genomics to cognitive performance genomics in quantified creativity and thinking fast and slow. Data Driven Wellness: From Self-Tracking to Behavior Change (2013)

[8] Tang, G., Pei, J., Luk, W.S.: Email mining: tasks, common techniques, and tools. Knowledge and Information Systems pp. 1–31 (2013)

[9] Zock, M.: Sorry, what was your name again, or how to overcome the tip-of-the tongue problem with the help of a computer? In: Proceedings of the 2002 workshop on Building and using semantic networks-Volume 11, pp. 1–6. Association for Computational Linguistics (2002)